什么是机器学习

> 机器学习： 让机器具有人一样的学习能力的技术，从堆积如山的数据中寻找出有用知识的数据挖掘技术。

一、学习种类

> 根据所处理的数据种类的不同，可以分为监督学习、无监督学习、强化学习。

- 监督学习：
  有求知欲的学生从老师那里获取知识、信息，老师提供对错提示、告知最终答案的学习过程。
  在机器学习里。学生对应于计算机、老师对应周围环境。根据学习过程中获得的经验、技能，对没有学习过的问题也可以做出正确解答，使计算机获得泛化能力是监督学习的最终目标。

  > 适用场景：手写文字识别、声音处理、图像处理、垃圾邮件分类与拦截、网页检索、基因诊断以及股票预测等。

  > 典型任务: 预测数值型数据的回归、预测分类标签的分类、预测顺序的排序。

- 无监督学习
  没有老师的情况下，学生自学的过程。不仅仅局限于解决像监督学习那样的有明确答案的问题，学习目标可以不必十分明确。

  > 适用场景：人造卫星故障诊断、视频分析、社交网站解析、声音信号解析、数据可视化、监督学习方法的前处理工具。

  > 典型任务： 聚类、异常检测。

- 强化学习

类似监督学习，目标一致，但是没有老师提示对错、告知最终答案的环节。如果在学习过程中不能从周围环境获取任何信息，它就变成了无监督学习。
它会自己对预测的结果进行评估，通过自我评估，为了得到最好的效果不断学习。

> 适用场景：机器人的自动控制、计算机游戏中的人工智能、市场战略的最优化

> 典型任务: 回归、分类、聚类、降维等。

二、典型任务

2.1 回归

把实函数在样本点附近加以近似的有监督的函数近似问题。

2.2 分类

对于指定的模式进行识别的有监督的模式识别问题。

分类问题中的输出样本不是具体的实数，而是分别代表类别的值。

> 所以也可以像回归问题那样，被看作函数近似问题。

> 对一个或者多个自变量和因变量之间的关系进行建模，求解的一种统计方法。

分类问题只是单纯对样本应该属于哪一类别进行预测，并根据预测结果准确与否来衡量泛化误差，这一点与回归不同。

2.3 异常检测

寻找输入样本中包含的异常数据。

如果已知正常数据和异常数据的例子，则与有监督的分类问题相同。
如果是无监督的异常检测，一般采用密度估计方法，把靠近密度中心的数据作为正常数据，偏离密度中心的数据作为异常数据。

2.4 聚类

同分类一样，也是模式识别问题，但是属于无监督学习。经常把“簇”代替“类别”。

2.5 降维

从高纬度数据中提取关键信息，将其转换为易于计算的低维度问题。

三、机器学习的方法

以分类问题为例，机器学习的主要流派有：产生式分类、判别式分类、频率派、贝叶斯派。

3.1 生成的分类和判别的分类

已知x的时候，如果能求得使分类类别y的条件概率$p(y|x)$达到最大值的类别$\hat{y}$的话，就可以进行模式识别了：

$\hat{y}$ = $argmax p(y|x)$

> argmax是取最大值时的参数。所以$max_y p(y|x)$是指当y取特定值时$p(y|x)$的最大值。而$argmax_y p(y|x)$是指当$p(y|x)$取最大值时对应的y的值。

条件概率$p(y|x)$通常叫做后验概率。应用训练集直接对后验概率进行学习的过程，叫做**判别式分类**.

通常又可以把后验概率$p(y|x)$表示为y的函数。

$$p(y|x) = p(x,y)/p(x) \propto p(x,y)$$

可以发现模式x和类别y的联合概率与后验概率是成比例的，所以可以通过让联合概率达到最大的方法使后验概率达到最大的类别$\hat{y}$。

在模式识别里，联合概率也称为数据生成概率，因此通过预测数据生成概率来进行模式识别的分类方法，称为生成的分类。

3.2 统计概率和朴素贝叶斯

在统计概率方法中，如何由训练集得到高精度的模式是主要的研究课题。
在朴素贝叶斯方法中，模式作为概率变量，对其先验概率加以考虑，计算与训练集相对应的后验概率，用贝叶斯定理，可以使用先验概率求解后验概率。

如果先验概率已知，后验概率可以按照贝叶斯定理进行精确计算，因此朴素贝叶斯算法中，如何精确地计算后验概率是一个主要的研究课题。